HW07 Application exercises: Egalitarianism and everything

Ellen Hsieh

```r
library(tidyverse)
library(ggplot2)
library(rsample)
library(broom)
library(rcfss)
library(glmnet)
library(pls)
library(caret)
library(earth)
library(iml)
library(patchwork)

options(digits = 3)
theme_set(theme_minimal())
set.seed(124)
```

```r
# load the data
gss_train <- read_csv("./data/gss_train.csv")
gss_test <- read_csv("./data/gss_test.csv")
```

1.

```r
# OLS
lm_cv <- train(egalit_scale ~ ., data = gss_train, method = "lm", metric = "RMSE",
               trControl = trainControl(method = "cv", number = 10), preProcess = c("zv"))

# ElasticNet
elastic_cv <- train(egalit_scale ~ ., data = gss_train, method = "glmnet", metric = "RMSE",
                 trControl = trainControl(method = "cv", number = 10),
                 preProcess = c("zv", "center", "scale"), tuneLength = 10)
# PCR
pcr_cv <- train(egalit_scale ~ ., data = gss_train, method = "pcr", metric = "RMSE",
                trControl = trainControl(method = "cv", number = 10),
                preProcess = c("zv", "center", "scale"), tuneLength = 20)

# PLS
pls_cv <- train(egalit_scale ~ ., data = gss_train, method = "pls", metric = "RMSE",
                trControl = trainControl(method = "cv", number = 10),
                preProcess = c("zv", "center", "scale"), tuneLength = 20)

# MARS
hyperparam_grid <- expand.grid( degree = 1:3, nprune = seq(2, 100, length.out = 10) %>% floor())
mars_cv <- train(egalit_scale ~ ., data = gss_train, method = "earth", metric = "RMSE",
              trControl = trainControl(method = "cv", number = 10),
              preProcess = c("zv"), tuneGrid = hyperparam_grid)
```

```r
# compare different models
summary(resamples(list(
  Linear_regression = lm_cv,
  PCR = pcr_cv,
  PLS = pls_cv,
  ElasticNet = elastic_cv,
  MARS = mars_cv
)))$statistics$RMSE %>%
  round(3) %>%
  kableExtra::kable() %>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"))
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Linear_regression	7.27	7.62	7.98	7.92	8.15	8.69	0
PCR	7.54	7.81	7.99	8.06	8.36	8.52	0
PLS	7.49	7.64	7.88	7.93	8.23	8.48	0
ElasticNet	7.03	7.48	7.80	7.76	8.06	8.43	0
MARS	7.25	7.58	7.71	7.74	7.99	8.24	0

As reported by the table above, MARS performs the best since it has the smallest RMSE among all models.

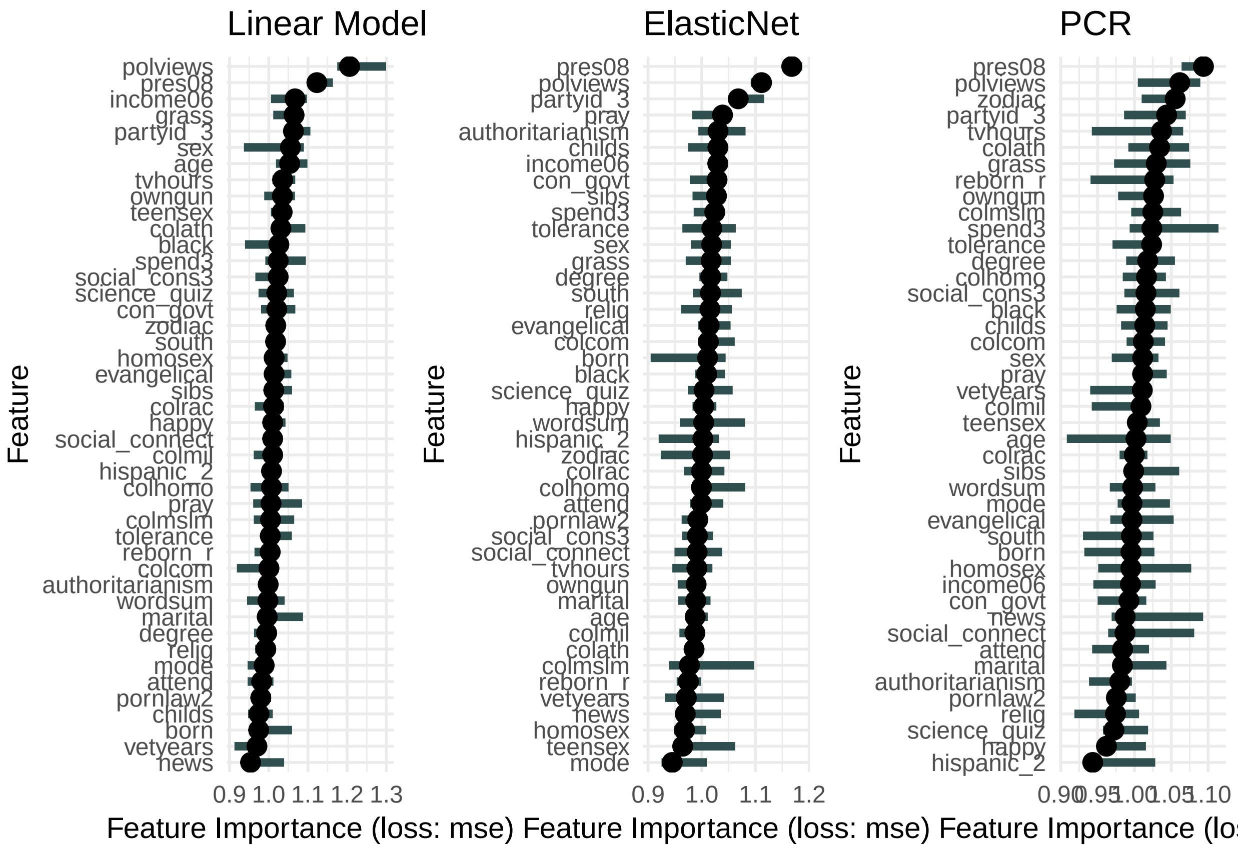
2.

```r
preds <- select(gss_train, -egalit_scale)
dep <- gss_train$egalit_scale

# Linear regression
lm_pred <- Predictor$new(model = lm_cv, data = preds, y = dep)

# ElasticNet
elastic_pred <- Predictor$new(model = elastic_cv, data = preds, y = dep)

# PCR
pcr_pred <- Predictor$new(model = pcr_cv, data = preds, y = dep)

# PLS
pls_pred <- Predictor$new(model = pls_cv, data = preds, y = dep)

# MARS
mars_pred <- Predictor$new(model = mars_cv, data = preds, y = dep)
```

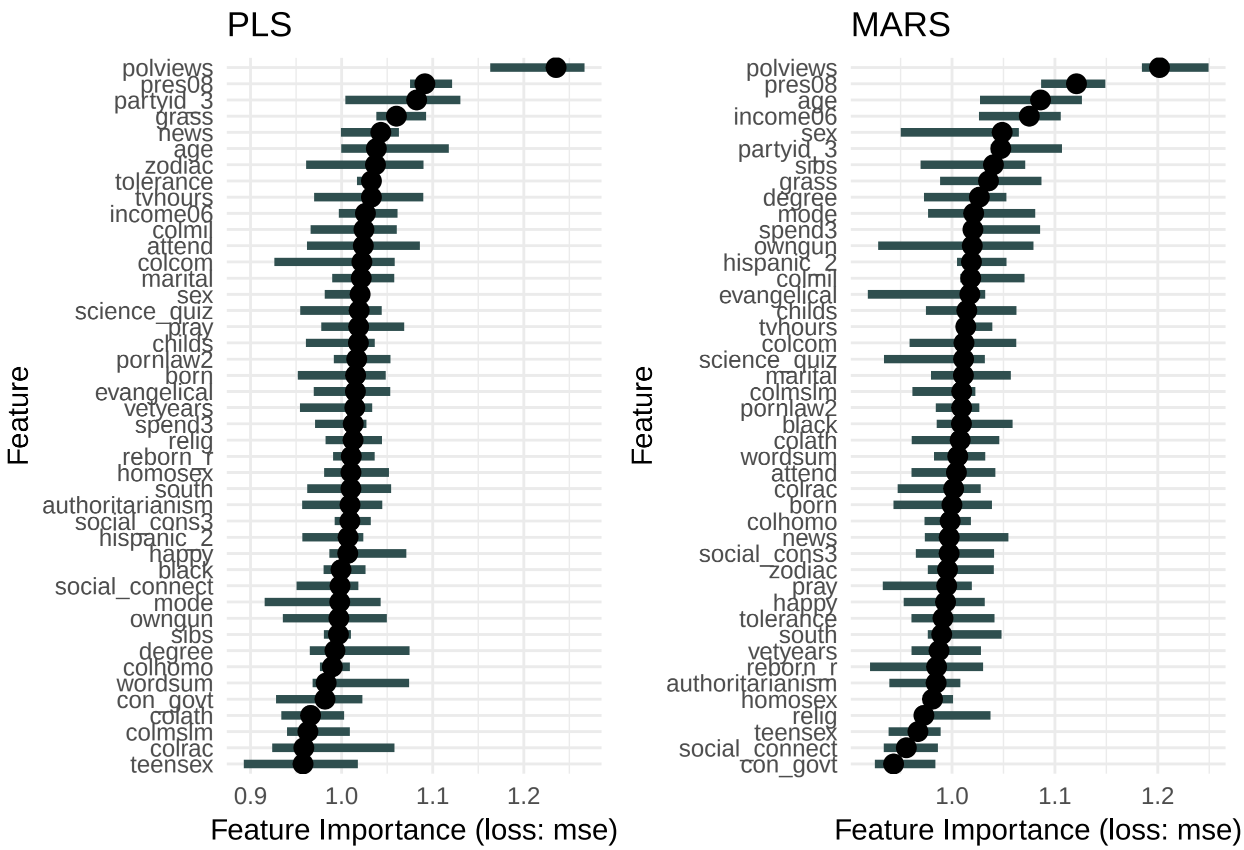
Feature Importance

```r
# get the feature importance from each model
imp_lm <- FeatureImp$new(lm_pred, loss = "mse")
imp_elastic <- FeatureImp$new(elastic_pred, loss = "mse")
imp_pcr <- FeatureImp$new(pcr_pred, loss = "mse")
imp_pls <- FeatureImp$new(pls_pred, loss = "mse")
imp_mars <- FeatureImp$new(mars_pred, loss = "mse")

# plot the feature importance
imp_lm_fig <- plot(imp_lm) + ggtitle("Linear Model")
```

```r
imp_elastic_fig <- plot(imp_elastic) + ggtitle("ElasticNet")
imp_pcr_fig <- plot(imp_pcr) + ggtitle("PCR")
imp_pls_fig <- plot(imp_pls) + ggtitle("PLS")
imp_mars_fig <- plot(imp_mars) + ggtitle("MARS")

imp_lm_fig + imp_elastic_fig + imp_pcr_fig
```




```r
imp_pls_fig + imp_mars_fig
```

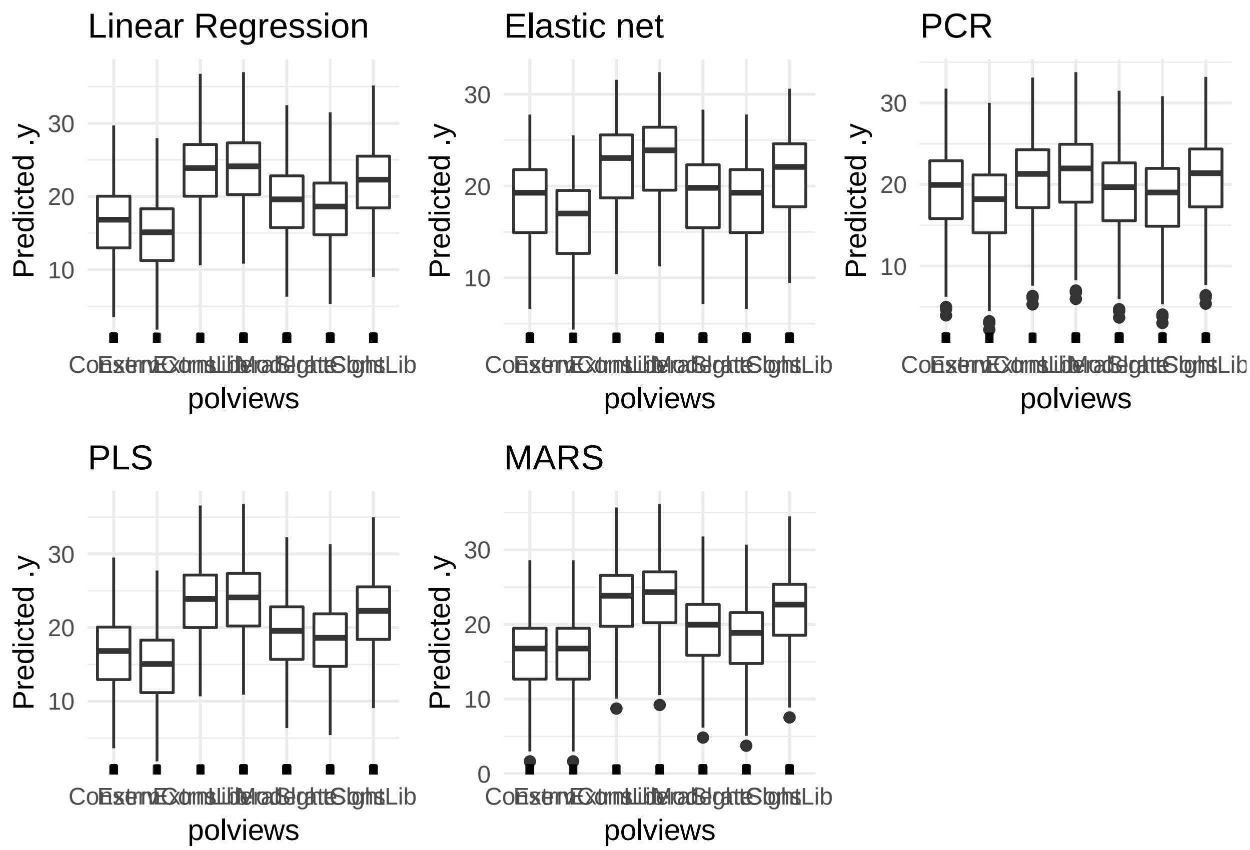


Avvording to the observation above, `polviews`, `pres08` are the most important features among all the models, and other important features are `partyid_3`, `income06` and `childs`.

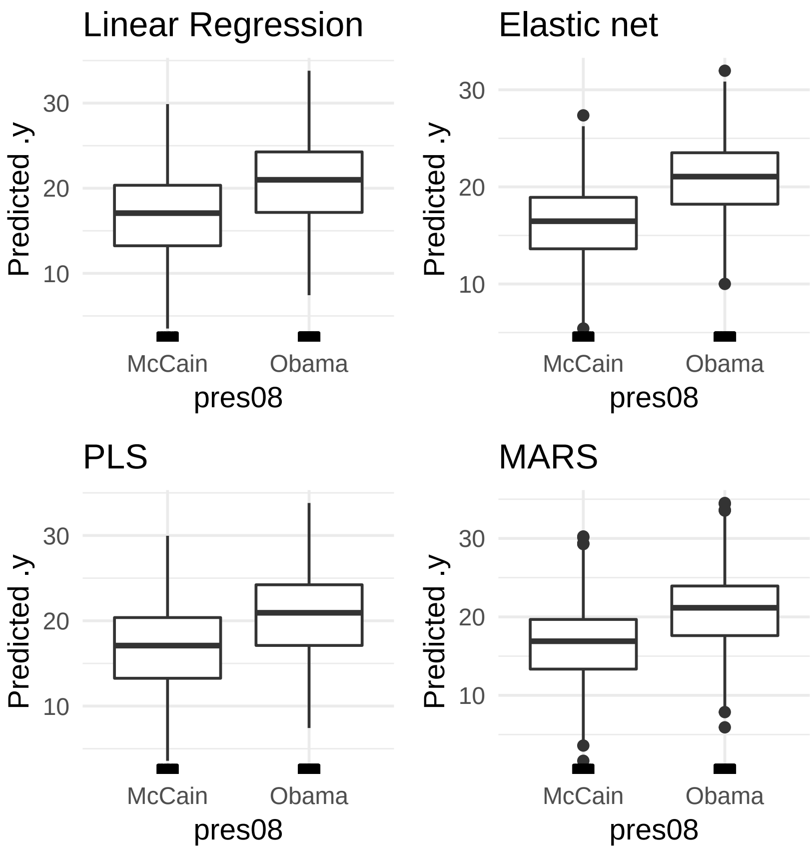
PDPs

```
predictors <- tibble(
  name = c("Linear Regression", "Elastic net", "PCR", "PLS", "MARS"),
  models = list(
    Linear= lm_pred,
    Elastic = elastic_pred,
    PCR = pcr_pred,
    PLS = pls_pred,
    MARS = mars_pred)
)

# get the PDPs and ICE for each important features
pdps <- predictors %>%
  mutate(polviews = map2(models, name, ~ FeatureEffect$new(.x, "polviews", method = "pdp+ice") %>%
                      plot() + ggtitle(.y)),

         pres08 = map2(models, name, ~ FeatureEffect$new(.x, "pres08", method = "pdp+ice") %>%
                      plot() + ggtitle(.y)),

         partyid_3 = map2(models, name, ~ FeatureEffect$new(.x, "partyid_3",method = "pdp+ice") %>%
                      plot() + ggtitle(.y)),

         income06 = map2(models, name, ~ FeatureEffect$new(.x, "income06", method = "pdp+ice") %>%
                      plot() + ggtitle(.y)),
```

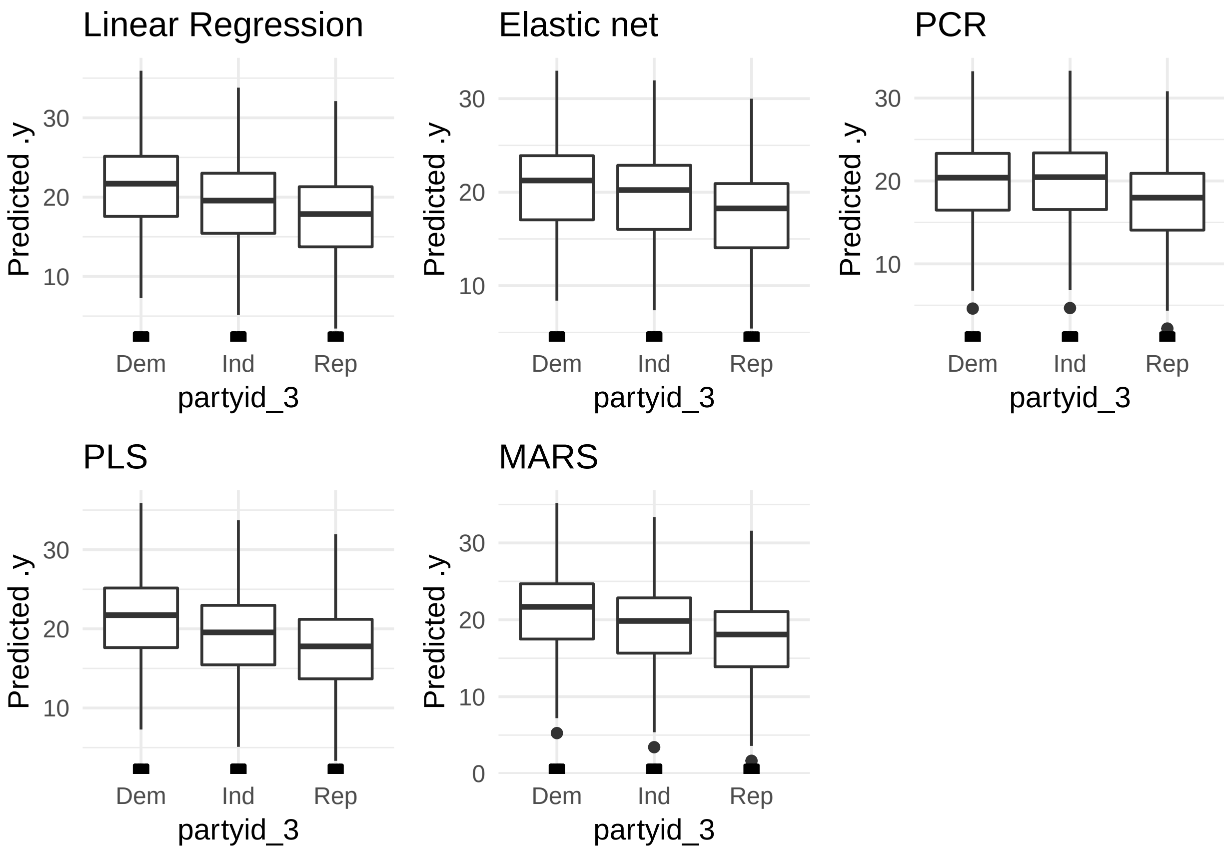
```
        childs = map2(models, name, ~ FeatureEffect$new(.x, "childs", method = "pdp+ice") %>%
                         plot() + ggtitle(.y)))

# plot the PDPS for the five most important features
wrap_plots(pdps$polviews)
```



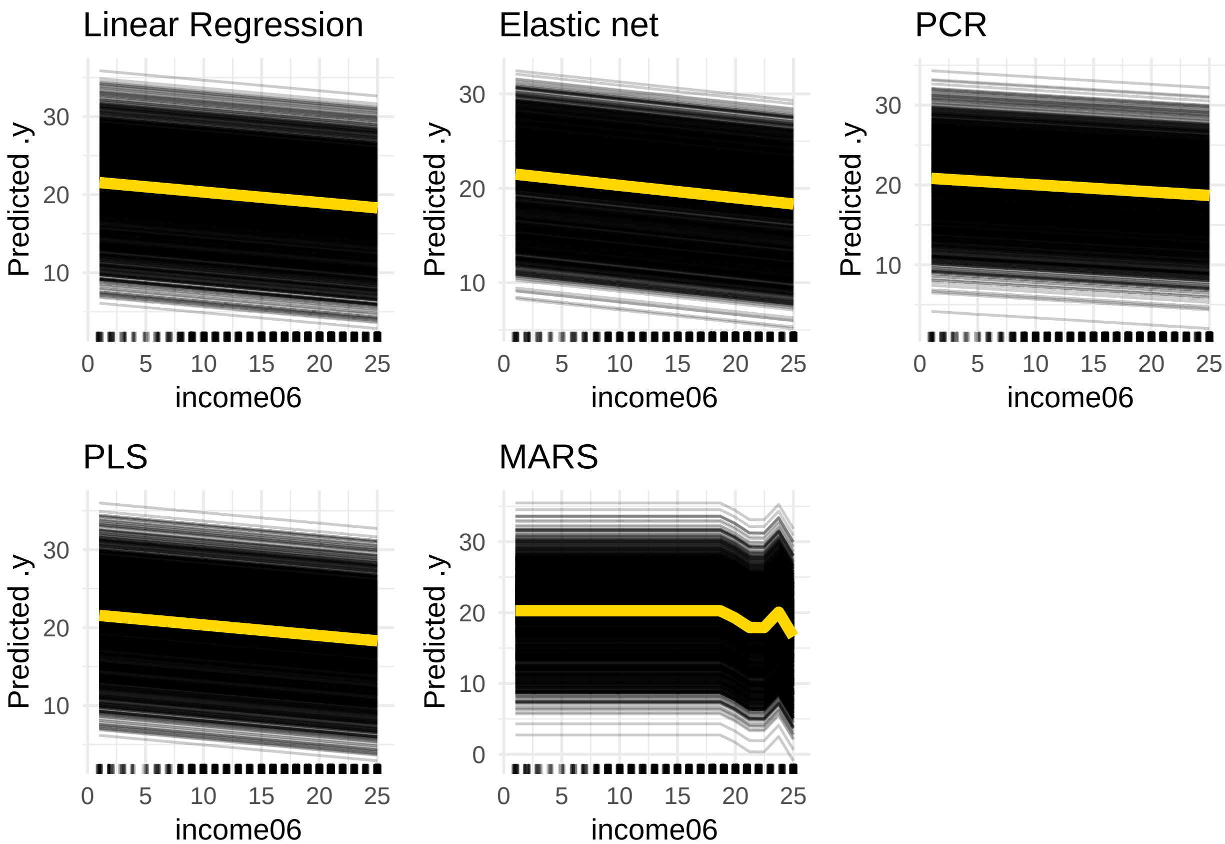

```
wrap_plots(pdps$pres08)
```



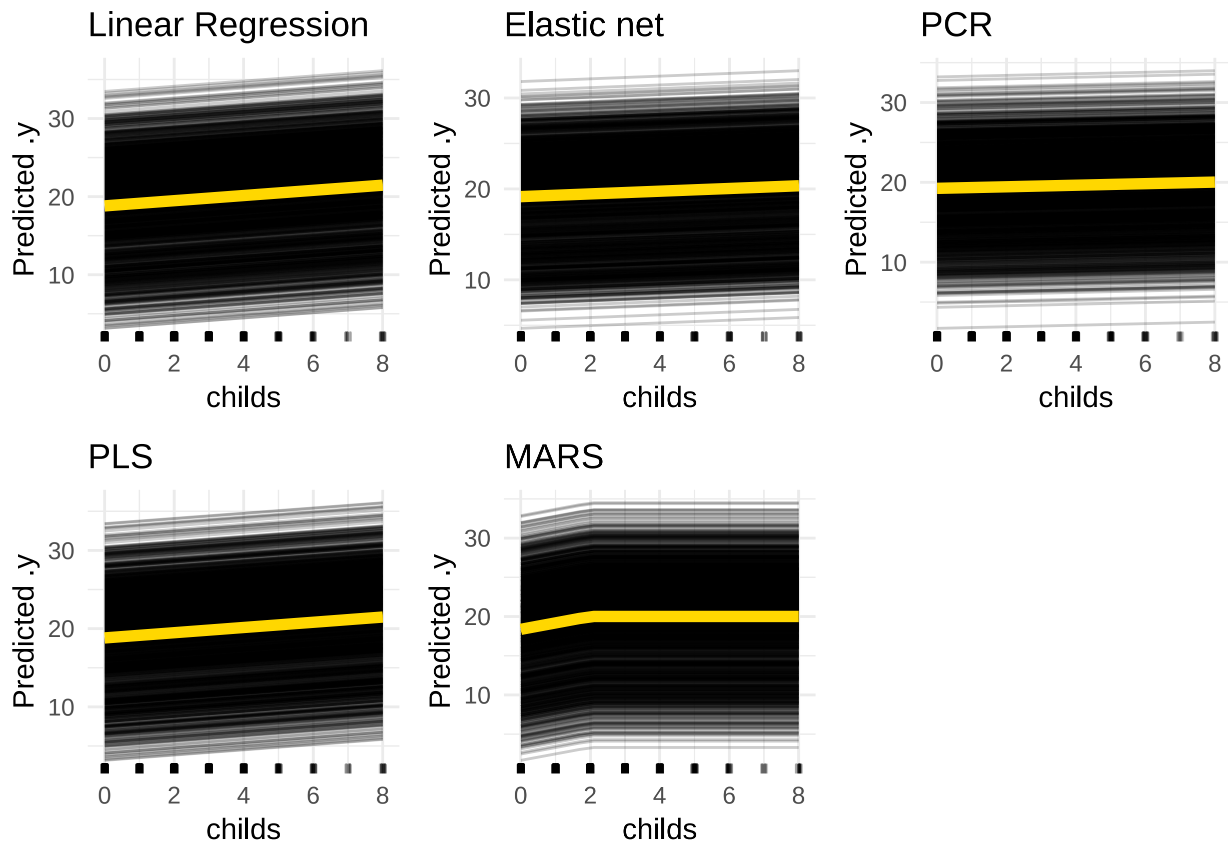

```
wrap_plots(pdps$partyid_3)
```




```r
wrap_plots(pdps$income06)
```




```
wrap_plots(pdps$childs)
```



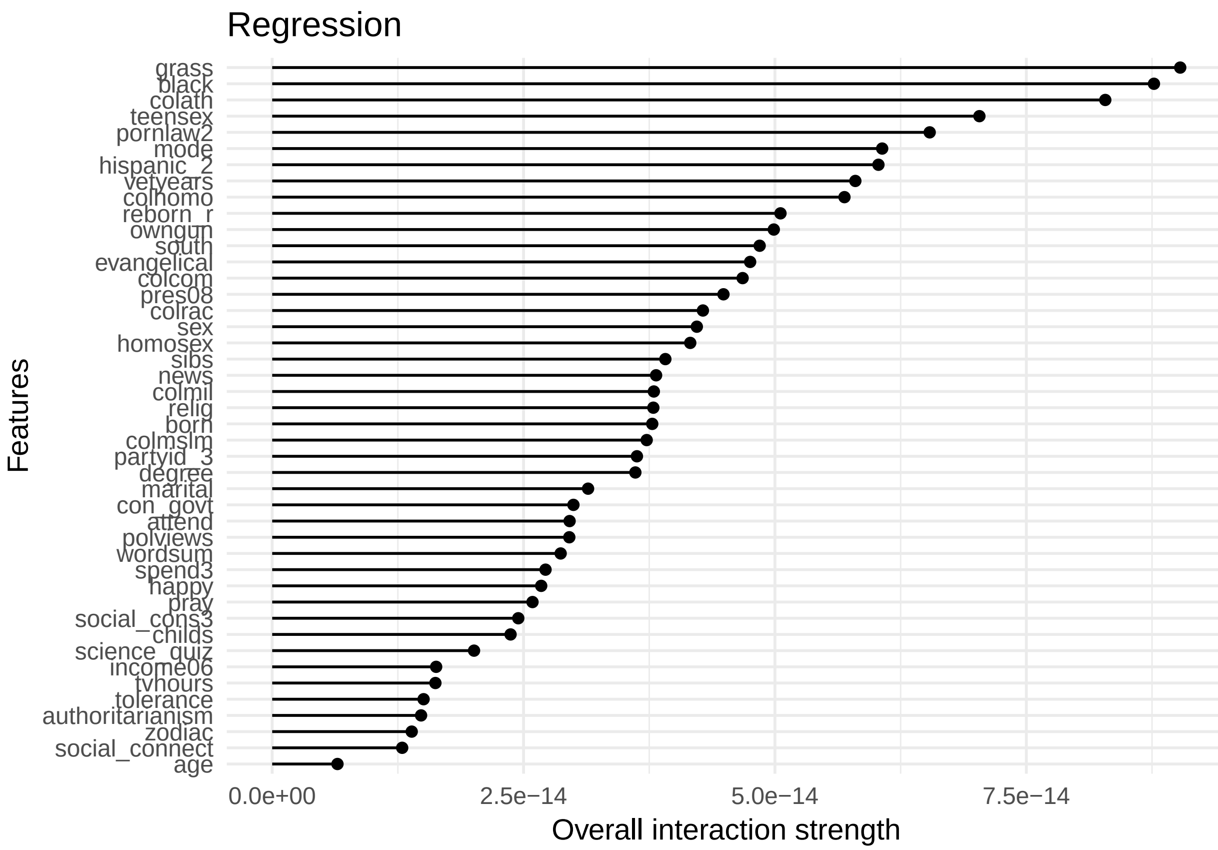
- As reported by the PDP of `polview`, the centrist are more egalitarian than those who have extreme political views.
- As reported by the PDP of `pres08`, those who voted for Obama are more egalitarian than those who voted for McCain.
- As reported by the PDP of `partyid_3`, the Democrats are more egalitarian than independents and Republicans. In addition, independents are more egalitarian than Republicans.
- As reported by the PDP of `income06`, people are less egalitarian when their level of income increases.
- As reported by the PDP of `childs`, people have more chidren are more egalitarian than others.

Feature Interaction

```
# linear model
lin_inter <- Interaction$new(lm_pred)
lin_inter_score <- lin_inter$results
lin_inter_score %>%
  arrange(-.interaction) %>%
  head(5)
```

```
##   .feature .interaction
## 1    grass     9.03e-14
## 2    black     8.77e-14
## 3   colath     8.29e-14
## 4  teensex     7.03e-14
## 5 pornlaw2     6.54e-14
```

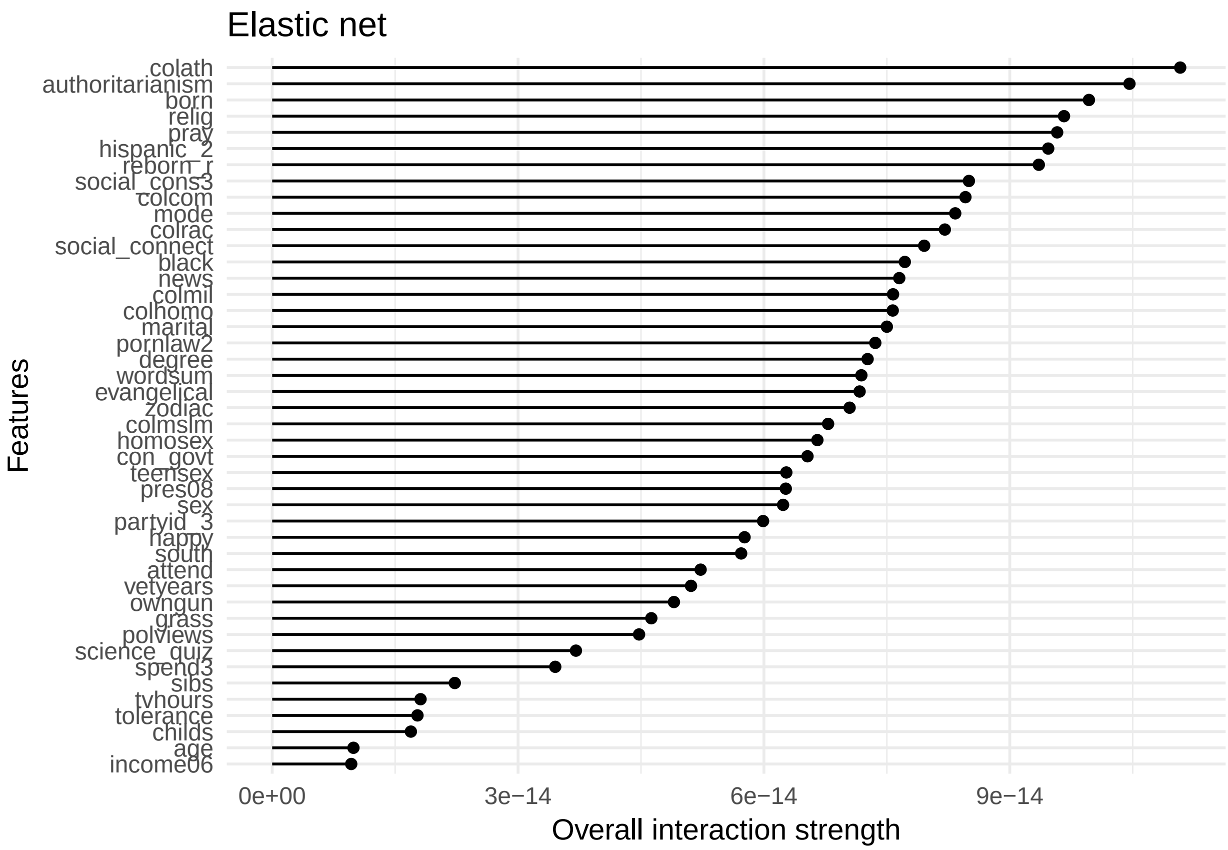
```
plot(lin_inter) + ggtitle("Regression")
```




```r
# ElasticNet
elas_inter <- Interaction$new(elastic_pred)
elas_inter_score <- elas_inter$results
elas_inter_score %>%
  arrange(-.interaction) %>%
  head(5)
```

```
##            .feature .interaction
## 1            colath     1.11e-13
## 2 authoritarianism     1.05e-13
## 3              born     9.97e-14
## 4             relig     9.66e-14
## 5              pray     9.58e-14
```

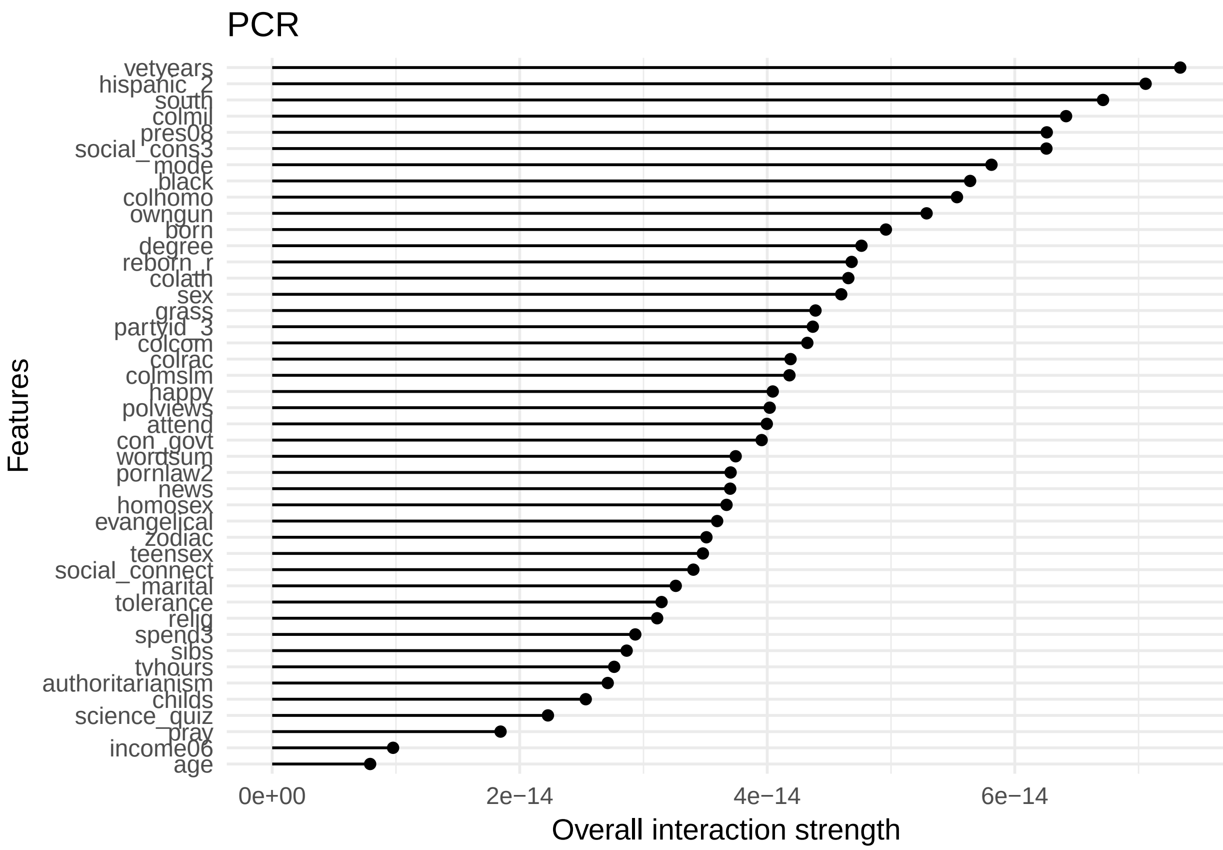
```r
plot(elas_inter) + ggtitle("Elastic net")
```




```r
# PCR
pcr_inter <- Interaction$new(pcr_pred)
pcr_inter_score <- pcr_inter$results
pcr_inter_score %>%
  arrange(-.interaction) %>%
  head(5)
```

```
##     .feature .interaction
## 1   vetyears     7.34e-14
## 2 hispanic_2     7.06e-14
## 3      south     6.71e-14
## 4     colmil     6.41e-14
## 5     pres08     6.26e-14
```

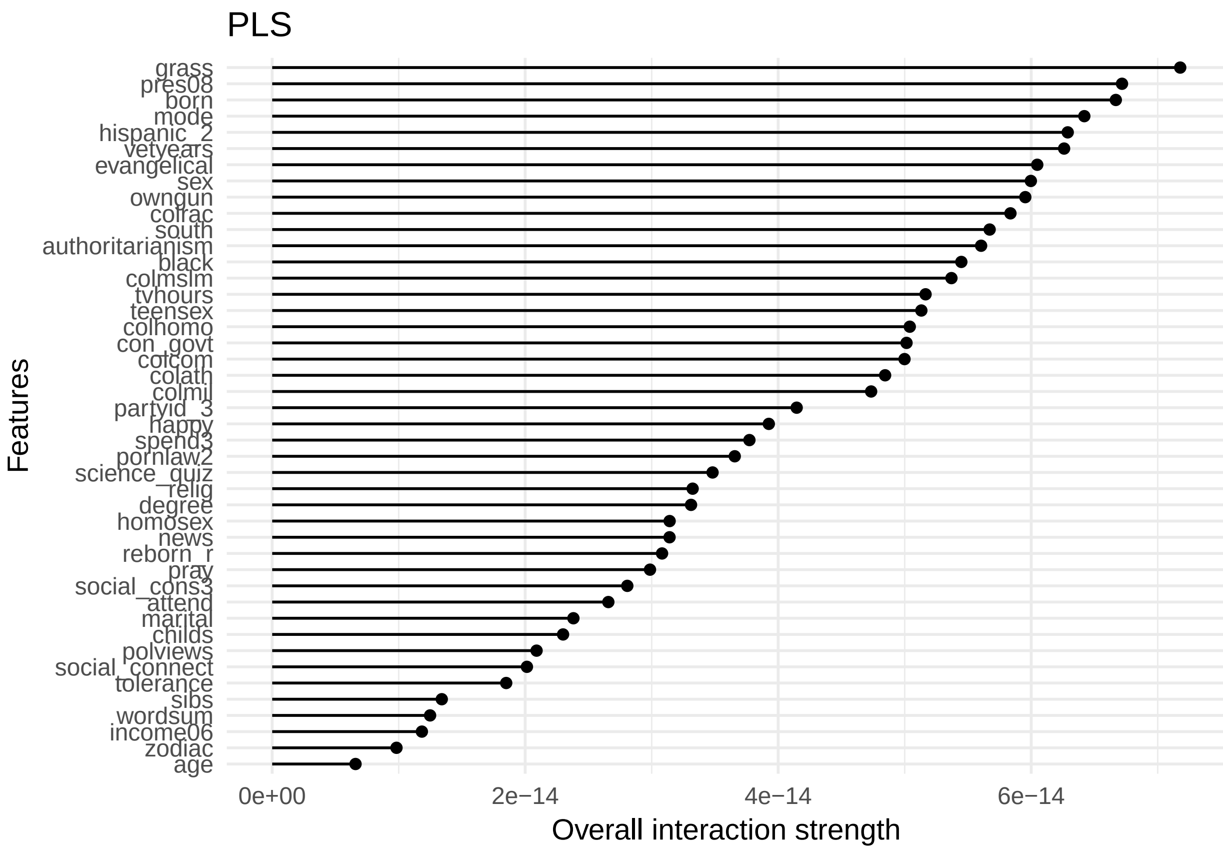
```r
plot(pcr_inter) + ggtitle("PCR")
```




```r
# PLS
pls_inter <- Interaction$new(pls_pred)
pls_inter_score <- pls_inter$results
pls_inter_score %>%
  arrange(-.interaction) %>%
  head(5)
```

```
##      .feature .interaction
## 1       grass     7.18e-14
## 2      pres08     6.72e-14
## 3        born     6.67e-14
## 4        mode     6.42e-14
## 5 hispanic_2     6.29e-14
```

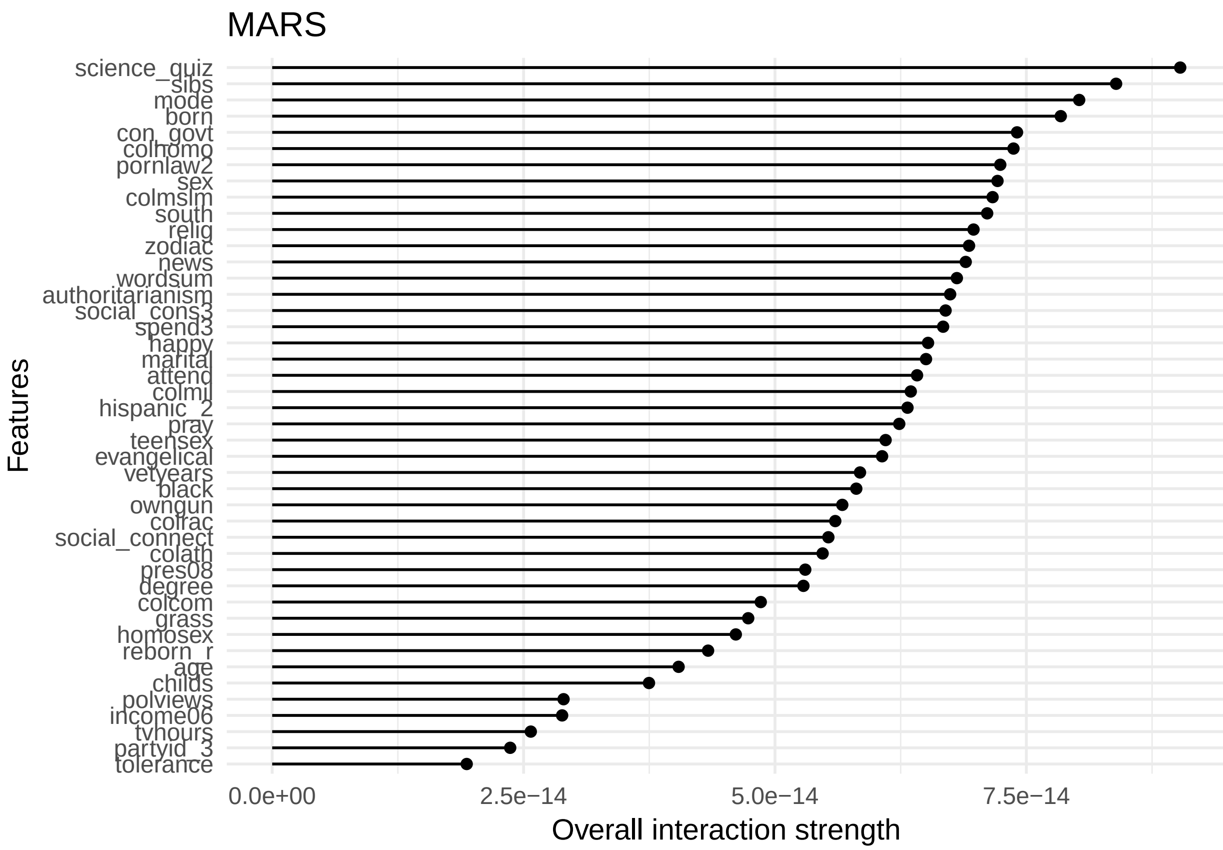
```r
plot(pls_inter) + ggtitle("PLS")
```




```r
# MARS
mars_inter <- Interaction$new(mars_pred)
mars_inter_score <- mars_inter$results
mars_inter_score %>%
  arrange(-.interaction) %>%
  head(5)
```

```
##        .feature .interaction
## 1 science_quiz     9.03e-14
## 2          sibs     8.39e-14
## 3          mode     8.03e-14
## 4          born     7.84e-14
## 5      con_govt     7.41e-14
```

```r
plot(mars_inter) + ggtitle("MARS")
```



3.

```r
# apply the test set to the optimal model
mars_cv$bestTune
```

```
##   nprune degree
## 3     23      1
```

```r
pred_test <- predict(mars_cv, newdata = gss_test)
actual <- gss_test$egalit_scale

# calculate the test MSE of the model
sse_test <- sum((pred_test - actual)**2)
mse_test <- sse_test / length(actual)
mse_test
```

```
## [1] 65.4
```

The optimal MARS model is with nprun = 23, degree = 1. This model generalizes well to the test set, since its MSE is similar to the training MSE.